ST 540 – Midterm Take-home Exam (Due 4/17)

Andrew Emerson (ajemerso)

1. Introduction

This report summarizes the findings from modeling a 50-year meteorological dataset. The dataset has 50 years of sea surface temperature (SST) readings for the 6 months leading up to hurricane season, taken at 10 distinct locations in the Atlantic Ocean. The objective is to determine the most predictive months and locations for the number of hurricanes that make landfall on the US Atlantic Coast. At the end of this report, I will use the best performing model to predict the number of hurricanes for the 50th (final) year.

2. Methods

I focus on three variations of the same general family of Bayesian models. Specifically, I use a **Poisson** likelihood for the response variable due to its ability to model count data (or events). The number of hurricanes making landfall is a discrete count, so it makes sense to use a distribution that will model the number of occurrences based on a yearly rate, $\boldsymbol{\lambda_i}$, where *i* represents the year. The general form of this Poisson regression can be stated as follows:

$$N_i \sim Poisson(\lambda_i),$$

$$\log(\lambda_i) = \alpha + \sum_{j=1}^{p} X_{ij}\beta_j$$

In this model, N_i corresponds to the number of hurricanes in year *i*, and *p* corresponds to the number of covariates (subject to change based on specifications of the model). We use a logarithmic link function to be able to exponentiate the linear predictors used in the model and ensure the rate is positive. In this report, I compare variations of this family of models, using different covariates and experimental setups. I choose the best performing model by comparing the DIC and WAIC scores for each model, and then I determine the most important covariates for making the ultimate prediction.

I maintain the same general priors for each model. For each β_j, I use a double exponential prior to shrink the parameter towards 0 if the covariate provides little value (as in Bayesian LASSO regression). I place an uninformative normal prior on α, and I use an uninformative gamma prior for the precision for each of the β_j's double exponential prior. The prior distributions used in this general framework are as follow:

$$\beta_j \sim DoubleExponential(0, \tau) \text{ for all covariates,}$$

$$\alpha \sim Normal(0, 0.01),$$

$$\tau \sim Gamma(0.1, 0.1)$$

The specifics for each model are as follow. In my **first model**, I treat each monthly SST reading for each location as a separate covariate. This means there are a total of $p = 60$ covariates used to describe the hurricane counts. In this naïve model, note that the number of covariates is larger than the number of observations (i.e., $p > n$), where $n = 49$. In my **second model**, I reduce the number of covariates by taking the mean of all 6 months leading up to hurricane season. This creates a total of 10 covariates, one for each location. Thus, each covariate will be the 6-month average SST temperature for a particular location. In this case, the number of covariates is much less than the number of observations ($p < n$). In my **third and final model**, I attempt to introduce a temporal component to the second model. Each covariate is defined as the difference between the 6th month's temperature at each location and the 6-month mean temperature for each location. There are still 10 covariates, but now we will be examining how predictive the 6th month is in relation to the other months, for each location.

3. Computation

For fair comparison between models, I used the same experimental setup throughout the analysis. Each model was constructed in JAGS, using a burn-in of 10,000 and then sampling 20,000 times. To evaluate convergence, I used two parallel chains and validated the effective size and Gelman-Rubin diagnostic. Additionally, I used a thinning parameter of 10.

4. Model Comparisons

To compare the models, I ensured each model converged (i.e., all parameters had a G-R diagnostic of ~1 and effective sample size > 1000). To choose the best model, I compare the DIC and WAIC values for each variation of the Poisson regression described previously. The model comparisons are shown below:

Model	*Mean Deviance*	*DIC Penalty*	*Penalized Deviance (DIC)*	*WAIC*	*P_w*
1	**255.8**	28.3	**284.1**	**287.8**	**24.0**
2	332.3	**10.1**	342.4	366.9	28.8
3	446.6	**10.1**	456.7	514.0	55.7

From the table, it is clear that model 1 performs the best. This is surprising, as model 2 and 3 have significantly less covariates and are much less complex. It is surprising, additionally, because model 1 is a very naïve representation of the data, while models 2 and 3 attempt to use domain intuition.

5. Results

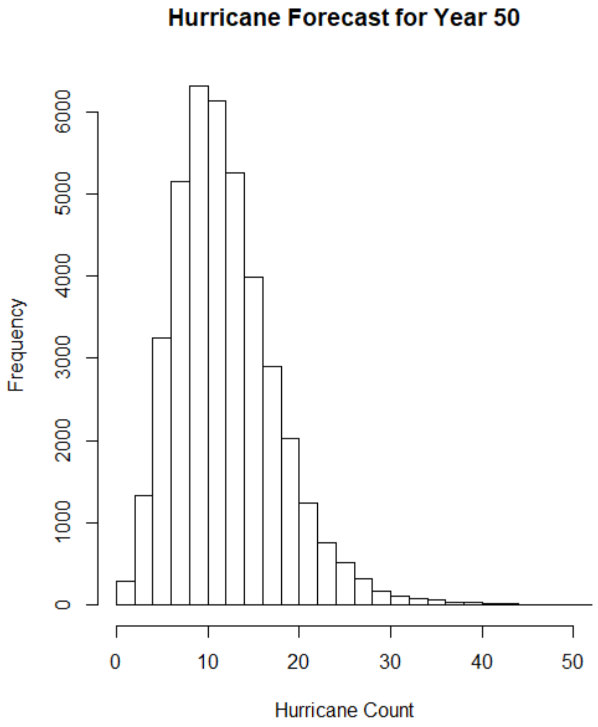
Using model 1, I report the significant findings regarding the locations and months that are more predictive of the number of hurricanes. I will not summarize all 60 covariates, but I will report a few of the most predictive. As a reminder, many of the parameters will be ~0 because of the double exponential priors. The most significant coefficients from my analysis using model 1 are summarized in the table below.

Month	*Location*	*2.5%*	*50%*	*97.5%*
4	10	0.0784	0.3169	0.5529
5	8	-0.0234	0.1409	0.3627
2	7	-0.0214	0.1326	0.3646
3	5	-0.0273	0.1119	0.3259
2	6	-0.0386	0.1023	0.3131
3	9	-0.3362	-0.0807	0.0529

It appears as if the middle months (2, 3, 4/5) were most predictive of the number of hurricanes. Specifically, the 2nd and 3rd months were positive predictors (i.e., the higher the SST in these months, the higher the hurricanes). Additionally, the locations in the northern part of the map (i.e., 6-10) were most predictive of the hurricane counts for each year. Since each covariate in model 1 represented both a location and month, the coefficients represent the joint effect of a particular month and location. To determine the most impactful months and locations, I report covariates whose 95% intervals did not contain 0 and covariates whose absolute mean values were relatively high.

6. Prediction

Using model 1 to predict the number of hurricanes for year 50, I used the learned (sampled) coefficients for each of the covariates and plugged in the 50th year SSTs. This was done via MCM, and the posterior distribution is shown below. The most likely value of this distribution is **10 hurricanes**, which is the prediction for the 50th year. The 95% credible interval for this distribution is [4, 26]. The 50th percentile is 12.



ST 540 – Midterm Take-home Exam (Due 4/17) JAGS Code

Andrew Emerson (ajemerso)

```
n <- length(Y) - 1  # 49 due to last observation being the prediction

burn <- 10000
iters <- 20000

model_string_1 <- textConnection("model{
# Likelihood
for(i in 1:n){
  Y[i] ~ dpois(lambda[i])
  log(lambda[i]) <- alpha + inprod(X[,,i],beta[])
}

# Priors
for(j in 1:60){
  beta[j] ~ ddexp(0, inv.var)
}
alpha ~ dnorm(0, 0.01)
inv.var ~ dgamma(0.1,0.1)

# Prediction
predict_dist ~ dpois(lambda_predict)
log(lambda_predict) <- alpha + inprod(X[,,50],beta[])

# WAIC
for(i in 1:n){
  like[i] <- dpois(Y[i], lambda[i])
}

}")

data <- list(Y=Y, X=X, n=n)
model_1 <- jags.model(model_string_1, data=data, n.chains=2, quiet=TRUE)
update(model_1, burn)
samples_1 <- coda.samples(model_1, variable.names=c("alpha", "beta", "like", "predict_dist"),
n.iter=iters, n.thin=10)
```